第 46 卷 第 8 期 计 算 机 工 程 2020 年 8 月
Vol. 46 No. 8 Computer Engineering August 2020

·开发研究与工程应用· 文章编号：1000-3428(2020)08-0297-08 文献标志码：A 中图分类号：TP391

融合 CRF 与规则的老挝语军事领域命名实体识别方法

何阳宇[1]，晏 雷[2]，易绵竹[1]，李宏欣[1,3]

（1. 中国人民解放军战略支援部队信息工程大学（洛阳校区），河南 洛阳 471003；
2. 昆明理工大学 信息工程与自动化学院，昆明 650500；3. 密码科学技术国家重点实验室，北京 100878）

摘 要：针对老挝语军事领域命名实体识别存在的规则制定不准确、覆盖不全等问题，提出一种融合条件随机场与规则的识别方法。通过分析老挝语语言和领域文本特点，选取词、词性、通名、指界词和词典等原子特征构建组合特征模板，在自建标注语料上训练条件随机场模型，并利用测试语料进行测试。为识别错例，加入能够表达语言确定性的规则进行后处理，以提升识别性能。实验结果表明，该方法总体准确率、召回率和 F 测度值分别达到 91.49%、90.96% 和 91.22%，可有效提高老挝语军事领域命名实体识别效果。

关键词：命名实体识别；军事领域；老挝语；条件随机场；信息抽取

开放科学（资源服务）标志码（OSID）：

中文引用格式：何阳宇，晏雷，易绵竹，等. 融合 CRF 与规则的老挝语军事领域命名实体识别方法[J]. 计算机工程，2020，46(8)：297-304.

英文引用格式：HE Yangyu，YAN Lei，YI Mianzhu，et al. Named entity recognition method for Laotian in military field combining CRF and rules[J]. Computer Engineering，2020，46(8)：297-304.

Named Entitiy Recognition Method for Laotian in Military Field Combining CRF and Rules

HE Yangyu[1]，YAN Lei[2]，YI Mianzhu[1]，LI Hongxin[1,3]

（1. PLA Strategic Support Force Information Engineering University（Luoyang Campus），Luoyang，Henan 471003，China；
2. College of Information Engineering and Automation，Kunming University of Science and Technology，Kunming 650500，China；
3. State Key Laboratory of Cryptology，Beijing 100878，China）

【Abstract】To address the problems of inaccurate formulation and incomplete coverage of existing methods for Laotian Named Entity Recognition（NER）in the military field，this paper proposes a method combining Conditional Random Field（CRF）and rules. By analyzing the characteristics of Laotian and domain texts，the method selects the atomic features such as the word，the part of speech，the general name，the boundary word and the dictionary to construct a combined feature template. The CRF model is trained on the self-built tagged corpus，and tested by using the test corpus. To identify wrong examples，it adds rules that can express language certainty for post-processing to improve recognition performance. Experimental results show that the final overall accuracy，recall rate and F measures of this method reach 91.49%，90.96% and 91.22% respectively，effectively improve the Laotian Named Entity Recognition（NER）in military field.

【Key words】Named Entity Recognition（NER）；military field；Laotian；Conditional Random Field（CRF）；Information Extraction（IE）

DOI：10.19678/j.issn.1000-3428.0055363

0 概述

随着大数据时代的到来，互联网已经成为军事情报获取的重要来源，但海量冗杂的数据也带来了“信息过载”的问题，命名实体识别（Named Entity Recognition，NER）是解决这一问题的有效手段。

基金项目：国家自然科学基金（61701539）；密码科学技术国家重点实验室开放课题（MMKFKT201825）；国防科技创新特区项目。

作者简介：何阳宇（1992—），男，博士研究生，主研方向为自然语言处理、知识图谱；晏 雷，硕士研究生；易绵竹，教授、博士生导师；李宏欣，讲师、博士。

收稿日期：2019-07-02 **修回日期：**2019-08-21 **E-mail：**mugedawen@163.com

“命名实体”是在第六届消息理解会议(MUC-6)上首次使用的,可以简单地定义为“任何一个可以被专有名称指代的事物”[1]。在这次会议上,命名实体识别也作为信息抽取(Information Extraction,IE)的子任务被提出[2],之后迅速成为大数据分析、文本意义理解、语义表示、知识管理等研究领域的关键技术之一。近年来兴起的以知识图谱为基础的智能检索,其核心单元即为实体。

命名实体识别即识别文本中的专有名称,并将其划分到预先定义的类别。按照技术手段可以分为基于规则的方法、基于统计的方法以及深度学习的方法[3]。最早出现的命名实体识别系统大多是基于规则的,从 20 世纪 90 年代开始,统计方法逐渐成为主流。常用的统计模型有支持向量机(Support Vector Machine,SVM)、最大熵模型(Maximum Entropy,ME)、隐马尔可夫模型(Hidden Markov Model,HMM)、条件随机场(Conditional Random Fields,CRF)等,这类模型通常将实体识别任务形式化为从文本输入到特定目标结构的预测,使用统计模型来建模输入与输出之间的关联,并使用机器学习方法来学习模型的参数。例如,隐马尔可夫模型将命名实体识别视为字符串分类问题[4],条件随机场模型则将实体识别转化为序列标注问题[5]。最近广受欢迎的深度学习也被应用到了命名实体识别任务中,目前主要的命名实体深度学习架构有两类[6]:一类是神经网络-条件随机场(NN-CRF)架构[7],在该架构中,卷积神经网络(CNN)和长短期记忆(LSTM)网络被用来学习每一个词位置处的向量表示,基于该向量表示,NN-CRF 解码该位置处的最佳标签;另一类是采用滑动窗口分类的思想使神经网络学习句子中每一个 N-gram 的表示,然后预测该 N-gram 是否为一个目标实体[8]。深度学习的方法虽然省去了统计方法中特征选取的过程,但需要更大规模的训练语料。对于老挝语这种低资源语言,构建大型标注语料库所需的人力和物力成本短期内是暂时无法承受的,将大量未标注或少量人工标注的数据集用于训练老挝语命名实体识别的统计模型更符合研究现状[9],并且实践证明单纯基于统计的方法会使状态搜索空间非常庞大,加入一定的规则等先验性知识也是有必要的。

关于老挝语命名实体识别的成果较少,老挝国内几乎没有专门研究,仅有文献[10]借助命名实体识别来提升老挝语分词的效果,其中的实体识别部分主要是利用规则的方法对人名和地名进行识别。国内主要有昆明理工大学对此进行了研究,其成果基本都采用机器学习的方法[11-12],取得了一定的成果,但囿于语料规模和质量,其识别结果难以泛化[13-14]。如果直接移植到军事领域,准确率可能会大幅下降。此外,尽管其中加入了一些规则和特征,但存在制定不准确、覆盖不全等问题,这势必也会影响识别效果。

本文提出一种融合 CRF 和规则的老挝语军事领域命名实体识别方法。首先在分析老挝语军事领域文本的基础上,选取了词、词性、指界词、通名和词典等特征训练得到 CRF 模型,从而实现老挝语军事领域命名实体的自动识别。然后对输出结果中的错例进行分析,并通过人工制定规则来提升识别性能。

1 老挝语军事领域命名实体识别类型

命名实体识别任务通常针对人名、地名和机构名三大类专有名词,结合具体研究领域和任务,本文需在此基础上进行增改。经综合考量,最终将老挝语军事领域命名实体分为人名(PER)、地名(LOC)、军事机构名(ORG)、武器装备名(WE)和军用设施名(FAC)等类型。

从广义上来讲,人名包括本名、别名、乳名、笔名、艺名等,但本文仅识别本名,即人的正式姓名称谓;地名是指某一特定空间位置上自然或人文地理实体的专有名称,自然地理实体包括山、河、湖、海、岛等,而人文地理实体包括国家、省、市、县、村等,即通常所说的行政地名;军事机构名可再分为指挥机构、编制单位、科研机构、军工企业、教育培训机构和医疗机构等几大类;武器装备名是武装力量用于实施和保障战斗行动的武器、武器系统和军事技术器材等的名称,包括枪械、火炮、坦克、装甲战斗车辆、作战飞机、战斗舰艇、弹药、导弹、水雷等战斗装备以及雷达、声呐、通信指挥器材、军用测绘器材、野战工程机械、军用车辆、保障舰船、辅助飞机、情报处理装备、电子对抗装备等保障装备[15];军事设施名是指用于军事目的的建筑、场地和设备等的专有名称,主要包括指挥工程、作战工程、军用机场、港口、码头、营区、训练场、试验场、军用洞库、仓库、军用通信、侦察、导航、观测台站和测量、导航、助航标志、军用公路、铁路专用线、军用通信输电线路、军用输油输水管道等。本文结合老挝实际情况,以军事工程、军事基地、军事交通设施、各类场地和塔台站为重点识别对象。

2 老挝语命名实体识别的难点

老挝语命名实体识别既有所有语言面临的共同难点,也具有其独特的个性难点,对此进行剖析有助于后续研究,具体如下:

1)英语等西方语言单词之间一般都有空格,并且专有名称首字母需大写,因此其实体边界非常易于确定,只需完成实体分类任务即可。而老挝语却不具备这样的先天优势,其缺乏丰富的词形变化和明显的形态标志,并且没有天然的词边界,分词、浅层句法分析等过程都会影响老挝语命名实体识别的效果。

2)丰富的语料资源对于命名实体识别任务来说相当重要,这也正是老挝语的不足。研究力量薄弱、关注度低等原因造成了可供老挝语命名实体识别使用的语料极为匮乏,专门针对军事领域的电子化资源则几乎没有,唯一的解决办法就是通过多渠道自行构建。即便如此,因为命名实体是一个相对开放的集合,新的命名实体会不断涌现,规模再大的语料库也难以做到及时更新和完全覆盖。

3)部分实体拼写较为随意,尤其是外来词,有时甚至不符合老挝语的拼写规则,老挝国内也没有权威机构对此进行规范管理。

4)老挝语命名实体常出现双语混合使用的情况,主要包括 2 种:第 1 种是完整的英语实体词汇出现在老挝语文本中,这就需要融入英语命名实体识别技术;第 2 种是老挝语和英语共同构成一个实体,如“ເຮືອບິນສູ້ຮົບ F22(F22 型战斗机)”。

5)老挝语中存在大量缩略词,这些缩略词往往就是命名实体,其形式较为多样,很难总结出规则,且有时会出现多个命名实体对应同一缩略词的现象。军事领域具有特殊性,文本表达通常言简意赅,因此缩略词的使用更为常见。

6)根据不同的任务和目的,实体的类别不再局限于人名、地名和机构名,出现越来越多的开放类别实体,本文所要识别的“武器装备名”等就属于此种情况。为了提高识别精度,便于后续应用,实体划分的颗粒度也越来越小,如本文中的地名可被细化为国家、省、市、县、村等。

7)自然语言处理的相关研究几乎都是完成某个层面的歧义消解,命名实体识别也不例外。其面临的歧义主要包括结构歧义和词义歧义两大类,结构性歧义一般是由连词造成的,在老挝语中主要有“ແລະ”和“ກັບ”等(老挝语中都是“和”“与”的意思)。比如,“ກະຊວງປກຊ ແລະ ປກສ(国防部和公安部)”,这个例子包含了两个部门,为了简洁,“公安部”省略了词头“ກະຊວງ(部)”,而“ກະຊວງແຜນການ ແລະ ການລົງທຶນ”则指的是一个部门——计划与投资部。词义性歧义还可分为“一对多”和“多对一”两种情况。所谓“一对多”是指同一名称可指向不同的实体,比如,“ຈຳປາ”既表示“占巴花”,也可作为人名,“ລາວ”既可指“老挝”这个国家,也有人称代词“他”的意思;而“多对一”指多个名称指向同一实体,这种情况往往由某个实体的别称、代称、简称等造成。比如,中国最新研制的“运-20”重型运输机,其代号为“鲲鹏”,如果“运-20”和“鲲鹏”都出现在军事文本中,一般可视为同一实体。

8)部分实体结构复杂,各个类别的实体互相嵌套。比如,“ປນຕຳຊັນ(汤姆森冲锋枪)”中的“ຕຳຊັນ”为人名,而这一实体整体为武器装备名。再如,“ກອງບັນຊາການທະຫານແຂວງອຸດົມໄຊ(万象省军事指挥部)”就是地名“ແຂວງອຸດົມໄຊ(万象省)”嵌套在机构名之中。这类结构较长且复杂的实体在识别时需要重点分析上下文和语法特征,准确定位实体边界。

3　基于 CRF 的老挝语军事领域命名实体识别

3.1　条件随机场

条件随机场自 2001 年 LAFFERTY 等人[5]提出以来,因其简单的操作原理和良好的性能在自然语言处理等领域迅速受到了广泛欢迎。之后,MCCALLUM[16]率先将其用于命名实体识别。经过不断改进,其成为目前命名实体识别中最成功的方法[17]。它是一种用于分割和标注序列数据的概率化结构模型,在已知观察序列 X 的情况下,计算输出标注序列 Y 的条件概率 $P(Y|X)$。

与隐马尔可夫模型(HMM)、最大熵马尔可夫模型(MEMM)等其他序列标注模型相比,CRF 弱化了独立性假设,只需考虑已经出现的观察序列的特性,能够充分利用上下文信息,易于融合不同的特征,同时其在全局范围内进行参数优化和解码,避免了 MEMM 和其他判别式马尔可夫模型会出现的标记偏置(Label Bias)问题。CRF 和 MEMM 之间的关键区别在于,MEMM 使用每个状态的指数模型来确定当前状态的下一个状态的条件概率,而 CRF 则利用单个指数模型来计算整个标注序列和给定观察序列的联合概率。因此,在不同状态下不同特征的权重可以相互替换[5]。

CRF 可以被视作一种无向图模型或者马尔可夫随机场[18]。从理论上来讲,只要在标注序列中表示一定的条件独立性,其图结构可以是任意的,但一般用来解决序列标注问题的是最为简单且常见的一阶链式结构,如图 1 所示。

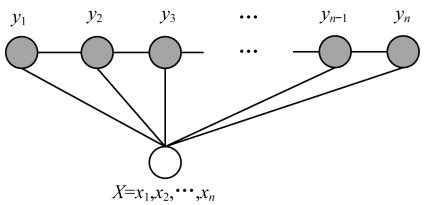

图 1　条件随机场链式结构

Fig. 1　Chain structure of CRF

本文定义:$X=x_1, x_2,\cdots,x_n$ 为给定的观察序列,即由 n 个词组成的老挝语语料,$Y=y_1,y_2,\cdots,y_n$ 为输出的标注序列,即为被预测出的实体标注序列。那么,输出序列的条件概率可定义为:

$$P(Y|X,\lambda)=\frac{1}{Z(X)}\exp\left(\sum_j \lambda_j t_j(y_{i-1},y_i,X,i)+\sum_k \mu_k s_k(y_i,X,i)\right) \tag{1}$$

其中,$Z(X)$为归一化因子,它可使所有可能状态序列概率之和为 1,可由式(2)得出,$t_j(y_{i-1},y_i,X,i)$为转移函数,表示对于观察序列 X 在当前位置 i 及前一位置 $i-1$ 上标注的转移概率,$s_k(y_i,X,i)$为状态函数,表示当前位置 i 的标注概率。以上两个函数统称为特征函数,都依赖于局部特征。在命名实体识别过程中,当满足特征模板条件时,取值为 1,否则取值为 0,λ_j和 μ_k分别为 t_j和 s_k对应的权值,可以通过最大似然函数在模型训练集上估算出来。

$$Z(X)=\sum_Y \exp\left(\sum_j \lambda_j t_j(y_{i-1},y_i,X,i)+\sum_k \mu_k s_k(y_i,X,i)\right) \tag{2}$$

在得出特征函数权值后,模型训练过程便基本完成。将观察序列 X(即测试语料)输入此模型,概率最大的命名实体标注序列 Y'便可通过维特比算法解码得出:

$$Y'=\arg\max P(Y\mid X,\lambda) \tag{3}$$

3.2 特征选取

在利用 CRF 模型进行命名实体识别的任务中,最为关键的一步就是构建与待识别对象相关联的特征模板,它直接影响识别系统的性能。

在制定特征模板时,需要先确定“观察窗口”,即当前位置词的前后 n 个词及其标注所构成的上下文语境。窗口大小的取值相当重要,开口过大会增加计算成本,影响模板通用性,出现过拟合现象;而开口太小则可能遗漏重要信息。结合老挝语军事领域文本特点和老挝语语言规律,本文将窗口大小设置为 5,即包括当前词及其前后各 2 个词。下面对所选取的各类原子特征进行阐释:

1)词特征:将词本身作为特征,记为 Fw_i($i=-2,-1,0,1,2$)(下同),w_0表示当前词,w_{-1}表示当前词左边第 1 个词,w_{-2}表示当前词左边第 2 个词,w_{+1}表示当前词右边第 1 个词,w_{+2}表示当前词右边第 2 个词。

2)词性特征:该特征也是与待标注词本身相关的特征,记为 Fp_i,p_i对应 w_i的词性。

3)通名特征:命名实体内部构件可分为通名、专名与饰名 3 种,其中通名是表示该命名实体类别属性的构件,如“ກົມລົດຕັ້ງ-ຍານເກາະ(坦克-装甲局)”中的“ກົມ(局)”即为机构通名。老挝语命名实体的出现多以通名开头,因此这是极具价值的实体识别启发信息。通名是一个相对封闭的集合,数量较为有限,本文将几个类型的实体通名进行归纳,共计 105 个,符号表示为 Ft_i。需要特别说明的是,汉语等语言中姓氏可以视作人名的通名,中国用姓的历史已有上千年,已经形成了数量较为固定的姓氏体系,完全可以采用穷尽的方法进行搜集。而老挝是在上世纪四十年代之后才逐渐形成用姓的习惯,据已知情况,老挝官方尚未对老挝的姓氏作过明确规定或统计,相关研究文献也不多。文献[19]对老挝人的取姓方式进行了总结,主要有“以长辈之名为姓”“以地名为姓”“从父之姓”“从夫之姓”“以职业为姓”与“自取之姓”等几大类,但仅仅是列举了少量已有的姓,并没有从大量统计结果中得出规律。因此,本文暂不考虑人名实体的通名,表 1 展示了部分通名列表。

表 1 老挝语命名实体通名(部分)

Table 1 General names of named entities in Laotian(partial)

实体类型	通名
LOC	ພູ(山)、ແມ່ນ້ຳ(河)、ແຂວງ(省)、ເມືອງ(县)、ບ້ານ(村)、ເຂດ(区)
ORG	ຄະນະ(委员会)、ກົມໃຫຍ່(总局)、ກອງບັນຊາການ(指挥部)、ກອງພົນ(师)
WE	ປືນ(枪械)、ລົດຕັ້ງ(坦克)、ເຮືອບິນ(飞机)、ລູກສອນໄฟ(导弹)、ລາດາ(雷达)
FAC	ຖານທັບ(基地)、ສະໜາມບິນ(机场)、ທ່າເຮືອ(港口)、ຄ້າຍທະຫານ(军营)

4)指界词特征:顾名思义,指界词就是对指示实体边界有重要作用的词,实体周围往往会有这类词的出现,可分为左指界词和右指界词,统一表示为 Fz_i。比如,句子“……ມີທ່ານພົນຕີອ່ອນສີ ແສນສຸກ ຮອງລັດຖະມົນຕີ ກະຊວງປ້ອງກັນປະເທດ……”中的职衔称谓“ພົນຕີ(少将)”和“ຮອງລັດຖະມົນຕີ(副部长)”可分别作为人名“ອ່ອນສີ ແສນສຸກ”的左右指界词。通过设计算法遍历标注语料库中所有实体左右各一个词,各类实体分别取频数前 15 的词作为指界词,然后加上人工搜集整理的部分词共同形成指界词集合。表 2 为部分指界词(示例中未将左右指界词分开列出,实际使用时须加以区分)。

表 2 老挝语实体识别指界词示例

Table 2 Examples of Laotian entity recognition boundary words

实体类型	指界词
PER	ພົນເອກ(上将)、ປອ/ດຣ(博士)、ສະຫາຍ(同志)、ໂດຍ(由)、ວ່າ(说)
LOC	ທີ່(关系代词)、ຢູ່(在……)、ໄປ(去)、ມາ(来)、ຮອດ(到)、ປະຈຳ(驻扎)
ORG	ຫົວໜ້າ(首长)、ແລະ(和)、ໃຫ້(使得)、ເປັນ(是、作为)、ດ້ວຍ(通过)
WE	ຍິງ(射击)、ຖິ້ມ(投掷)、ມອບ(交付)、ສົ່ງ(发送)、ຕິດຕັ້ງ(部署)
FAC	ຢູ່(在……)、ໃນ(在……里)、ຂອງ(的)、ທາງ(在……方面)、ໂจมตี(攻击)

5)词典特征:表示为 Fd_i,英、汉语等的命名实体识别研究已证明这一特征具备高度预测能力,尤其是针对地名等规模相对稳定的实体类型非常有效。本

文主要通过以下 3 个渠道构建常见词词典,即现有资源、网络爬虫和基于 Word2vec 的相似词推荐。

现有资源包括《老汉-汉老军事词典》[20]和老挝国家统计局出版的《2017 统计年鉴》[21]等。前者是目前国内唯一的老挝语军事领域词典,共收录 1.4 万余词条;后者对老挝主要山脉、河流以及省、县、村等行政单位进行了统计。

网路爬虫主要针对老挝人民军官网、老挝国防部官网以及维基百科老挝语版涉及军事的页面信息。前两个网站内有专门介绍军队组织架构等方面的内容,可直接提取存入词典。对于维基百科的爬取,只针对页面标题,因为维基百科每一个页面几乎都是对该页面标题的解释,而每一个标题多数情况下都代表一个实体。

基于 Word2vec 的相似词推荐的原理是训练语料生成词向量(Word Embedding)文件,然后以向量间的余弦距离(Cosine Distance)度量词语之间的相似度。可利用现有资源和网络爬虫获得的实体作为种子词集进行相似词推荐,为保证质量,将推荐阈值设定为 5,其余过程不再赘述。

将以上 3 个渠道获取的实体词条汇总后,根据准确性、广泛性和相关性原则,需要删除重复项、非名词词语、名词化的动词和形容词以及领域相关性较低的词,最终形成的常见词词典共包含 5 134 个实体词条。

根据上述各原子特征,本文依次进行组合叠加,构成如表 3 所示的特征模板。

表 3　老挝语军事领域命名实体识别组合特征模板

Table 3　Combined feature templates of named entity recognition in Laotian military field

序号	特征	说明
1	Fw_i	考虑当前词与前后两个词
2	Fw_i, Fp_i	考虑当前词与前后两个词及其词性
3	Fw_i, Fp_i, Ft_i	考虑当前词与前后两个词及其词性 + 通名
4	Fw_i, Fp_i, Ft_i, Fz_i	考虑当前词与前后两个词及其词性 + 通名 + 指界词
5	$Fw_i, Fp_i, Ft_i, Fz_i, Fd_i$	考虑当前词与前后两个词及其词性 + 通名 + 指界词 + 词典

4　实验结果与分析

4.1　语料获取及处理

老挝语目前还没有公开的命名实体标注语料,本文实验所采用的语料均为精通老挝语人士手工构建,主要来源为老挝人民军、老挝国防部等官方网站以及老挝通讯社 KPL、ABClaosnews 等老挝语主流网站的军事类新闻,语料规模约为 22.5M。将这些语料进行分词和词性标注等预处理后,由人工按照表 4 所示方法进行实体标注,然后使用标签集 BISO 对实体进行编码表示,B 表示实体的首词部分,I 表示实体的非首词部分,S 表示单个词构成的实体,O 表示非实体。本文 5 个类型实体对应的标签分别为 $\{B_{PER}, I_{PER}, S_{PER}, B_{LOC}, I_{LOC}, S_{LOC}, B_{ORG}, I_{ORG}, S_{ORG}, B_{WE}, I_{WE}, S_{WE}, B_{FAC}, I_{FAC}, S_{FAC}, O\}$。最终经过处理获得实验所需语料,其中,4/5 作为训练语料,1/5 作为测试语料。

表 4　老挝语命名实体人工标注示例

Table 4　Examples of named entities manual labeling in Laotian

实体类型	标注方法示例
PER	…… <ບຸນເຮືອງ ຈັນທະວົງ> [PER] ……
LOC	…… <ເມືອງຈັນທະບູລີ> [LOC] ……
ORG	…… <ກົມນະໂຍບາຍ> [ORG] ……
WE	…… <ເຮືອບິນສູ້ຮົບອາຍພົ່ນປະເພດມິກ> [WE] ……
FAC	…… <ສະໜາມພຶກແອບກອງພັນໃຫຍ່ 279> [FAC] ……

4.2　结果分析

为综合评价系统性能,模型训练完成后,将准确率(P)、召回率(R)以及 F 测度值(F-measure)作为评价指标进行测试,具体定义分别为:

$$P = \frac{\text{正确识别出的实体数}}{\text{所有识别出的实体数}} \times 100\% \tag{4}$$

$$R = \frac{\text{正确识别出的实体数}}{\text{所有应被识别的实体数}} \times 100\% \tag{5}$$

$$\text{F-measure} = \frac{2PR}{P+R} \times 100\% \tag{6}$$

本文针对不同的组合特征进行了 5 组实验,以对比各个特征对识别结果的影响,如表 5 ~ 表 9 所示。

表 5　基于组合特征 1 的识别结果(实验 1)

Table 5　Recognition results based on combined feature 1 (experiment 1)　%

实体类型	准确率	召回率	F 测度值
PER	91.01	92.08	91.54
LOC	82.73	85.11	83.90
ORG	80.09	79.65	79.87
WE	74.01	73.08	73.54
FAC	70.45	71.49	70.97
总体	81.42	82.82	82.11

表 6　基于组合特征 2 的识别结果(实验 2)

Table 6　Recognition results based on combined feature 2 (experiment 2)　%

实体类型	准确率	召回率	F 测度值
PER	92.84	94.24	93.53
LOC	93.00	93.03	93.01
ORG	82.24	82.11	82.17
WE	76.80	75.91	76.35
FAC	71.95	73.07	72.51
总体	86.20	85.88	86.04

表 7　基于组合特征 3 的识别结果(实验 3)
Table 7　Recognition results based on combined feature 3 (experiment 3)
%

实体类型	准确率	召回率	F 测度值
PER	94.33	95.76	95.04
LOC	93.51	93.60	93.55
ORG	84.18	83.91	84.04
WE	80.63	79.56	80.09
FAC	74.46	75.33	74.89
总体	87.66	87.30	87.48

表 8　基于组合特征 4 的识别结果(实验 4)
Table 8　Recognition results based on combined feature 4 (experiment 4)
%

实体类型	准确率	召回率	F 测度值
PER	94.53	95.80	95.16
LOC	93.77	93.88	93.82
ORG	84.26	83.99	84.12
WE	80.56	79.56	80.06
FAC	74.10	75.02	74.56
总体	87.77	87.42	87.59

表 9　基于组合特征 5 的识别结果(实验 5)
Table 9　Recognition results based on combined feature 5 (experiment 5)
%

实体类型	准确率	召回率	F 测度值
PER	94.69	95.86	95.27
LOC	94.73	94.77	94.75
ORG	84.65	84.31	84.48
WE	80.96	79.73	80.34
FAC	74.16	75.25	74.70
总体	88.38	88.00	88.19

通过观察表 5 ~ 表 9,从总体结果看,单独加入词特征已经达到了较好的识别效果,再加入词性特征后,3 项指标均又提升 4 个百分点左右,说明了词和词性特征对实体识别都十分重要。继续加入通名特征,指标提升放缓,大致提高了 1 个 ~ 2 个百分点,这可能有两个方面的原因:一是通名归纳不全,二是部分出现在语料中的实体没有加通名。比如,“ແຂວງຫລວງພະບາງ(琅勃拉邦省)”可写为“ຫລວງພະບາງ(琅勃拉邦)”。在前 3 个特征的基础上,加入指界词特征对效果的提高并不显著,各项指标仅有不到 1 个百分点的上升,最后加入词典特征同样如此。这可能是因为本文实验所用语料和词典规模较小,导致提取的指界词和常用词数量较为有限,无法涵盖可能出现的各种情况。具体到各类实体来看,人名和地名识别效果相对较好,评价指标几乎都在 90% 以上,而机构名、武器装备名和军事设施名的识别效果则与前两种实体相比有较大差距,其中原因为这 3 类实体具有简称多、嵌套多、歧义多等特点,现有的特征无法完全应对这些多变的语言现象。

4.3　基于规则的后处理

完成基于 CRF 模型的实体识别之后,本文对错误识别结果进行了分析,尝试加入适当的先验性知识,即能够表达语言确定性的规则,以期能够进一步提升系统性能。部分规则描述如下:

1)人名规则:造成人名错误识别的原因是当上下文无明显特征时,将临近的词作为人名的一部分或者将人名的一部分归入其他词,这可以通过词长 L_{PER}(即音节数量)规则来处理。本文以随机搜集的 500 个老挝人名为样本,进行音节数量的分布统计,如图 2 所示。从图 2 可以看出,老挝人名词长一般介于 3 ~ 8 之间,其中以 5 和 6 居多,占样本总数的 84.6%,因此可制定人名识别规则为 $3 \leqslant L_{PER} \leqslant 8$。

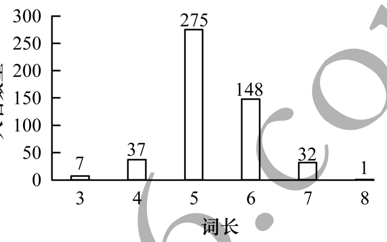

图 2　老挝人名词长分布
Fig. 2　Word length distribution of Lao name

2)地名规则:地名错例中较为常见的是上文提到的通名省略现象,因此可以规定词典中所有地名无通名与有通名等价;另一个错例类型为多级地名识别不全,比如,将“ບ້ານຫ້ວຍລ້ອງເມືອງຫົງສາ(洪洒县怀庄村)”识别为“ບ້ານຫ້ວຍລ້ອງ(怀庄村)”和“ເມືອງຫົງສາ(洪洒县)”两个地名。针对这一情况,由于老挝语中多级地名排列顺序是由小到大的,因此可以制定地名通名等级表,当地名出现时向后搜索直到无更高级别的地名出现为止。

3)机构名规则:嵌套其他类型的实体是机构名错例的主要原因之一,其中尤以人名和地名嵌套居多。通过观察语料,可以尝试规定当同一小句的机构名后面出现人名、地名时,只要中间不出现介词和动词,便可将人名、地名作为机构名的一部分。此外,当出现连接符号“-”时,如“ກົມລົດຕັງ-ຍານເກາະ(坦克-装甲局)”,将“-”后面的一个词也归入机构名。

4)武器装备名规则:武器装备名绝大部分是以其型号结尾,而型号主要由大写英文字母、阿拉伯数字、罗马数字和符号“ - ”“ + ”等要素构成,同时规定有无通名等价。

5)军事设施名规则:军事设施名面临的识别难点同样是人名和地名的嵌套现象,因此可采用与机构名类似的规则。

规则制定完成后,选择上述识别结果最好的实验 5 作为基础,利用规则进行后处理,结果如表 10 所示。与前 5 个实验的总体识别结果对比如图 3 所示。

表 10　加入规则后的识别结果(实验 6)

Table 10　Recognition results after adding rules (experiment 6)　%

命名实体类型	准确率	召回率	F 测度值
PER	96.84	96.73	96.78
LOC	96.55	95.90	96.22
ORG	89.74	88.42	89.08
WE	84.35	85.89	85.11
FAC	79.03	80.27	79.65
总体	91.49	90.96	91.22

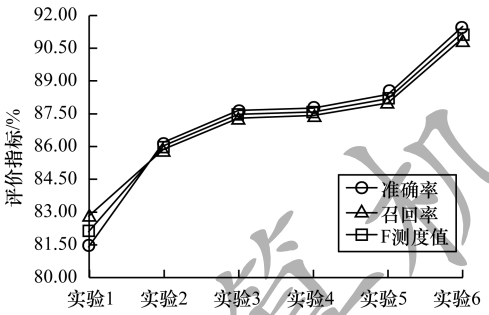

图 3　老挝语命名实体识别总体结果变化趋势

Fig. 3　Change trend of overall results of Laotian named entity recognition

可以看出,加入规则后的系统实体识别能力有了较为显著的提升,其中机构名、武器装备名和军事设施名的准确率、召回率和 F 测度值均提高了 4 个到 5 个百分点。由此证明了融合 CRF 和规则的方法具有可行性和有效性,可以在一定程度上弥补 CRF 模型的不足。

5　结束语

本文采用融合 CRF 和规则的方法对老挝语军事领域命名实体识别进行了研究。通过分析领域实体特点,选取词、词性、通名、指界词和词典等特征进行组合作为 CRF 模型的特征模板,利用测试语料进行测试,并对测试结果进行错例分析,人工制定具有针对性的规则进行后处理,进一步提升识别效果。实验结果表明,该选取特征可有效解决老挝语军事领域命名实体识别问题。由于目前没有公开的老挝语大型实体标注语料库,本文所用语料库为自行构建并且初次使用,语料的规模和质量还需进一步加强,下一步将尝试引入迁移学习技术[22]和自学习技术[23]来解决老挝语资源缺乏的现状,同时将对军事文件名、军事活动名等更多类别的军事领域实体识别进行研究。

参考文献

[1] JURAFSKY D, MARTIN J H. Speech and language processing [M]. 2nd ed. FENG Zhiwei, SUN Le, Translated. Beijing: Publishing House of Electronics Industry, 2018. (in Chinese)
JURAFSKY D, MARTIN J H. 自然语言处理综论[M]. 2 版. 冯志伟,孙乐,译. 北京:电子工业出版社,2018.

[2] HE Yangyu, YI Mianzhu, JIA Huixin, et al. A survey of Lao named entity recognition [J]. Modern Linguistics, 2018, 6(3): 449-461. (in Chinese)
何阳宇,易绵竹,贾惠心,等. 老挝语命名实体识别研究综述[J]. 现代语言学,2018,6(3):449-461.

[3] MAIHEMUTI Maimaiti, KAHAERJIANG Abiderexiti, AISHAN Wumaier, et al. Uyghur location names recognition based on conditional random fields and rules[J]. Journal of Chinese Information Processing, 2017, 31(6): 110-118. (in Chinese)
买合木提·买买提,卡哈尔江·阿比的热西提,艾山·吾买尔,等. CRF 与规则相结合的维吾尔文地名识别研究[J]. 中文信息学报,2017,31(6):110-118.

[4] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name [J]. Machine Learning, 1999, 34(1/2/3): 211-231.

[5] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the 18th International Conference on Machine Learning. New York, USA: ACM Press, 2001: 282-289.

[6] Special Committee on Language and Knowledge Computing of the Chinese Information Society of China. Knowledge graph development report (2018) [EB/OL]. (2018-08-18)[2019-02-15]. http://cips-upload.bj.bcebos.com/KGDevReport2018.pdf. (in Chinese)
中国中文信息学会语言与知识计算专委会. 知识图谱发展报告(2018)[EB/OL]. (2018-08-18)[2019-02-15]. http://cips-upload.bj.bcebos.com/KGDevReport2018.pdf.

[7] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]// Proceedings of NAACL-HLT'16. Sacramento, USA: Association for Computational Linguistics, 2016: 260-270.

[8] XU M, JIANG H, WATCHARAWITTAYAKUL S. A local detection approach for named entity recognition and mention detection[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: [s.n.], 2017: 1237-1247.

[9] WANG Hongbin, GAO Hongkui, SHEN Qiang, et al. Thai language names, place names and organization names entity recognition [J]. Journal of System Simulation, 2019, 31(5): 1010-1018. (in Chinese)
王红斌,郜洪奎,沈强,等. 泰语人名、地名、机构名实体识别研究[J]. 系统仿真学报,2019,31(5):1010-1018.

[10] SRITHIRATH A, SERESANGTAKUL P. A hybrid approach to Lao word segmentation using longest syllable level matching with named entities recognition [C]//Proceedings of International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology. Krabi, Thailand: [s.n.], 2013: 1-5.

[11] YANG Mengjie. Research on Lao language named entity recognition method[D]. Kunming: Kunming University of Science and Technology, 2016. (in Chinese)
杨梦杰. 老挝语命名实体识别方法的研究[D]. 昆明: 昆明理工大学, 2016.

[12] DUAN Shaopeng. Research on Lao language entity recognition [D]. Kunming: Kunming University of Science and Technology, 2017. (in Chinese)
段韶鹏. 老挝语命名实体识别研究[D]. 昆明: 昆明理工大学, 2017.

[13] HAN Rui. Research on Chinese-Lao language bilingual named entity recognition and alignment method[D]. Kunming: Kunming University of Science and Technology, 2018. (in Chinese)
韩锐. 汉老双语命名实体识别及对齐方法研究[D]. 昆明: 昆明理工大学, 2018.

[14] HUANG Yuquan. Research on identification of Lao language names, places and institutions[D]. Kunming: Kunming University of Science and Technology, 2018. (in Chinese)
黄于权. 老挝语人名、地名及机构名识别研究[D]. 昆明: 昆明理工大学, 2018.

[15] LEI Shujie, XING Fukun. Types and patterns of the English names of military weapons and equipments[J]. China Terminology, 2019, 21(1): 14-20. (in Chinese)
雷树杰, 邢富坤. 英文武器装备名的构成类型与构造模式研究[J]. 中国科技术语, 2019, 21(1): 14-20.

[16] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of the 7th Conference on Natural Language Learning. Edmonton, Canada: [s. n.], 2003: 188-191.

[17] ZONG Chengqing. Statistical natural language processing[M]. 2nd ed. Beijing: Tsinghua University Press, 2013. (in Chinese)
宗成庆. 统计自然语言处理[M]. 2 版. 北京: 清华大学出版社, 2013.

[18] WALLACH H M. Conditional random fields: an introduction[EB/OL]. (2004-02-24)[2019-04-26]. http://repository.upenn.edu/cis_reports/22.

[19] SU Tingting. A study of the name culture of Lao Lao people[D]. Nanning: Guangxi University for Nationalities, 2015. (in Chinese)
苏婷婷. 老挝佬族姓名文化研究[D]. 南宁: 广西民族大学, 2015.

[20] HUANG Yong, QIN Hailun. Lao-Chinese language Chinese-Lao language military dictionary[M]. Beijing: Yi Wen Publishing Military, 2009. (in Chinese)
黄勇, 覃海伦. 老汉-汉老军事词典[M]. 北京: 军事谊文出版社, 2009.

[21] Lao National Bureau of Statistics. 2017 statistical yearbook[M]. Vientiane, Lao: Lao National Bureau of Statistics, 2018. (in Chinese)
老挝国家统计局. 2017 统计年鉴[M]. 万象, 老挝: 老挝国家统计局, 2018.

[22] WU Hui, LÜ Li, YU Bihui. Chinese named entity recognition based on transfer learning and BiLSTM-CRF[J]. Journal of Chinese Computer Systems, 2019, 40(6): 1142-1147. (in Chinese)
武惠, 吕立, 于碧辉. 基于迁移学习和 BiLSTM-CRF 的中文命名实体识别[J]. 小型微型计算机系统, 2019, 40(6): 1142-1147.

[23] ZHONG Zhinong, LIU Fangchi, WU Ye, et al. Chinese named entity recognition combined active learning with self-training [J]. Journal of National University of Defense Technology, 2014, 36(4): 82-88. (in Chinese)
钟志农, 刘方驰, 吴烨, 等. 主动学习与自学习的中文命名实体识别[J]. 国防科技大学学报, 2014, 36(4): 82-88.

编辑 索书志